Discrete mutations in colorectal cancer correlate with defined microbial communities in the tumor microenvironment

Authors: Michael. B. Burns[1,2,#,*], Emmanuel Montassier[3,4], Juan Abrahante[5], Timothy K. Starr[1,6,7] Dan Knights[4], Ran Blekhman[1,2,*]

Affiliations:

[1]Department of Genetics, Cell Biology, and Development, University of Minnesota, Minneapolis, MN, 55455; USA.
[2]Department of Ecology, Evolution, and Behavior, University of Minnesota, Saint Paul, MN, 55108; USA.
[3]Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN, 55455; USA.
[4]Université de Nantes, EA 3826 Thérapeutiques cliniques et expérimentales des infections. Faculté de médecine, 1 Rue G Veil, 44000 Nantes, France.
[5]University of Minnesota Informatics Institute, University of Minnesota, Minneapolis, MN, 55455; USA.
[6]Masonic Cancer Center, University of Minnesota, Minneapolis, MN, 55455; USA.
[7]Department of Obstetrics, Gynecology and Women's Health, University of Minnesota, Minneapolis, MN, 55455; USA.

*Contact Information:
Michael B. Burns; michaelburns@umn.edu
Ran Blekhman; blekhman@umn.edu

[#]Current affiliation:
Department of Biology, Loyola University Chicago, Chicago, IL, 60626; USA.

Abstract

Variation in the gut microbiome has been linked to colorectal cancer (CRC), as well as to host genetics. However, we do not know whether genetic mutations in CRC tumors interact with the structure and composition of the microbial communities surrounding the tumors, and if so, whether changes in the microbiome can be used as a predictor for tumor mutational status. Here, we characterized the association between CRC tumor mutational landscape and its proximal microbial communities by performing whole-exome sequencing and microbiome profiling in tumors and normal colorectal tissue samples from the same patient. We find a significant association between loss-of-function mutations in relevant tumor genes and pathways and shifts in the abundances of specific sets of bacterial taxa. In addition, by constructing a risk index classifier from these sets of microbes, we accurately predict the existence of loss-of-function mutations in cancer-related genes and pathways, including MAPK and Wnt signaling, solely based on the composition of the microbiota. These results can serve as a starting point for understanding the interactions between host genetic alterations and proximal microbial communities in CRC, as well as for the development of individualized microbiota-targeted therapies.

Introduction

The human gut is host to approximately a thousand different microbial species consisting of both commensal and potentially pathogenic members[1]. In the context of colorectal cancer (CRC), it is clear that bacteria in the microbiome play a role in human cell signaling[2–11]; for example, in the case of CRC tumors that are host to the bacterium *Fusobacterium nucleatum*, the microbial genome encodes a virulence factor, *FadA*, that can activate the β-catenin pathway[12]. In addition, several attempts have been made to predict CRC status using the microbiome as a biomarker[13–16]. It has been shown that focusing on a single bacterial species, *F. nucleatum*, it is possible to predict some clinically relevant features of the tumor present[17]. However, as only a minority of CRCs are host to *F. nucleatum*, this is a somewhat limited application[18]. Other specific microbes have been linked to CRC, including *Escherichia coli* harboring polyketide synthase (pks) islands, as reported by one group[19,20] and enterotoxigenic *Bacteroides Fragilis* (ETBF) by another[21–23]. The mechanism of action of these associations is still under investigation with *F. nucleatum* being the most clearly developed[12].

In healthy individuals, host genetic variation can affect the composition of the microbiome[24–29], and the associated human genetic variants are enriched in cancer-related genes and pathways[25]. However, it is still unknown whether somatic mutations in host cells can affect the composition of the microbiome that directly interacts with host tissues. Here, we aim to find (i) whether variation in somatic mutational profiles in CRC tumors is associated with variation in the microbiome; (ii) which host genes and bacterial taxa drive the association; (iii) how these patterns can shed light on the molecular mechanisms controlling host-microbiome interaction in the tumor microenvironment; and (iv) whether this correlation can be used to construct a microbiome-based predictor of genes and pathways mutated in the tumor.

Results

Changes in the microbiome reflect tumor stage.

We performed whole-exome sequencing on a set of 88 samples, comprised of 44 pairs of tumor (adenocarcinomas) and normal colon tissue sample from the same patient, with previously characterized tissue-associated microbiomes[2]. The mutations in each of the tumors' protein-coding regions were identified relative to the patient-matched normal sample and annotated as either synonymous, non-synonymous, or loss-of-function (LoF) mutations (Supplementary Figs. 1-2, and Supplementary Tables 1-2). The mutations were collapsed by gene as well as by pathways using both Kyoto Encyclopedia of Genes and Genomes (KEGG) and pathway interaction database (PID) annotations[30–33].

We first investigated the relationship between microbial communities and tumor stage (Fig. 1). We hypothesize that the structure and composition of the associated microbiome can be affected by relevant physiological and anatomical differences between the tumors at different stages that would provide different microenvironmental niches for microbes. We identified the changes in the microbial communities surrounding each tumor as a function of stage by grouping the tumors into low stage (stages 1-2) and high stage (stages 3-4) classes and applied linear

discriminant analysis (LDA) effect size (LEfSe) to the raw operational taxonomic unit (OTU) tables corresponding to these tumors (Supplementary Tables 3-4)[34]. The set of taxon abundances was transformed to generate a single value representing a risk index classifier for membership in the low-stage or high-stage group (Fig. 1a; see Methods). To ascertain the error associated with these risk indices, a leave-one-out (LOO) cross-validation approach was applied. We also used the LOO results to generate receiver operating characteristic (ROC) curves and to calculate the area under the curve (AUC; see Fig. 1b). In addition, we performed a permutation test to assess the method's robustness (Supplementary Table 4). Using this approach, we demonstrate that the changes in abundances of 31 microbial taxa can be used to generate a classifier that distinguishes between low-stage and high-stage tumors at a fixed specificity of 80% and an accuracy of 77.5% (P = 0.02 by Mann-Whitney U test, and P = 0.007 by a permutation test; Supplementary Table 4).The resulting changes seen in our analysis of the microbial communities that vary by tumor stage were similar to those found in previous studies, including one using a Chinese patient cohort[4,35]. In both cases, there were significant changes among several taxa within the phylum *Bacteroidetes*, including *Porphyromonadaceae*, *Paludibacter*, and *Cyclobacteriaceae* (Fig. 1 and Supplementary Table 4).

Tumor mutations correlate with consistent changes in the proximal microbiome.

Next, we attempted to use a similar approach to classify tumors based on mutational profiles. We initially focused on individual genes that harbor loss-of-function (LoF) mutations, as those, we predicted, would be the most likely to have a physiologically relevant interaction with the surrounding microbiome. A prevalence filter was applied to include only those mutations that were present in at least 10 or more patients at the gene level. The raw OTU table was collapsed to the level of genus for the analysis. A visualization of the correlations between gene-level mutational status and the associated microbial abundances revealed differing patterns of abundances that suggests an interaction between the 11 most prevalent LoF tumor mutations and the microbiome (Supplementary Fig. 3). We hypothesized that the presence of mutation-specific patterns of microbial abundances could be statistically described by prediction of tumor LoF mutations in individual genes using the microbiome. For each of eleven genes that passed prevalence filtering cutoff, we identified the associated microbial taxa (Fig. 2a and Supplementary Tables 5-6), generated risk indices for each patient (Fig. 2b-c), and plotted the mean differences in abundances for a subset of microbial taxa interacting with each mutation (Fig. 2d). We found that we are able to use microbiome composition profiles to predict the existence of tumor LoF mutations in the human genes *APC*, *ANKRD36C*, *CTBP2*, *KMT2C*, and *ZNF717* (Q-value = 0.0011, 0.0011, 0.019, 0.019, and 0.055, respectively, by permutation test after False Discovery Rate (FDR) correction for multiple tests with a Q value threshold of 0.10; Fig. 2). The risk indices for each mutation were generated using sets of microbial taxa that ranges from 22 (*ZNF717*) to 53 (*ANKRD36C*) taxa (Supplementary Table 5). The taxa that showed the most dramatic differences in abundance when comparing tumors with and without mutations are shown in Fig. 2d. For example, the abundance of *Christensenellaceae* is relatively lower in tumors with *APC* mutations, but relatively higher in tumors with *ZNF717* mutations.

Next, we applied our interaction prediction approach, as described above, to the pathway-level mutational data (see Methods). Following visualization of the pathway level abundances (Supplementary Figs. 4-5) and applying the model, we found that each of the 21 KEGG pathways passing prevalence filter were able to be significantly predicted with a fixed specificity of 80% and an accuracy up to 86% (Q-values < 0.02 by permutation test after FDR correction; Fig. 3a-d, Supplementary Figs. 6-7, and Supplementary Table 7), as were 15 of the 19 tested PID pathways (Q-values < 0.04 by permutation test after FDR correction) (Fig. 3e-h, and Supplementary Figs. 8-9, and Supplementary Table 7). The taxon abundances that were specifically associated (direct or inverse correlations) with each of the LoF mutations in the genes and pathways can be found in Supplementary Tables 8-11 and Supplementary Fig. 10. In general, the number of taxa within each of the sets used to generate the risk indices was lower than that used for the gene-level analyses (average of 37 taxa per gene-associated set compared to 7 taxa per set associated with mutations in KEGG or PID pathways). When comparing results using the gene-level interactions and the pathway level interactions, for instance looking at mutations in *APC* (Fig. 2) and comparing them to mutations in the KEGG-defined Wnt signaling pathway and the PID-defined Canonical Wnt signaling pathway (Fig. 3), the interactions at the pathway level are more statistically significant (AUC for *APC* = 0.81, KEGG = 0.88, PID = 0.90). This trend is consistent and can be visualized as a density histogram of interaction prediction accuracies (Supplementary Fig. 11).

Predicted microbiome interaction network affected by tumor mutational profile

Lastly, we assessed the correlations between taxa among tumors with and without LoF mutations (Fig. 4; see methods). We found striking differences in structure of the network comparing tumors with and without a Lof mutation in *APC* the correlations between taxa (Fig. 4a). For example, in tumors with mutations in *APC*, the abundance of *Christensenellaceae* is positively correlated with *Rhodocyclaceae* and negatively correlated with *Pedobacter*. In tumors lacking LoF mutations in *APC*, these correlations are lost and *Christensenellaceae* is instead negatively correlated with *Saprospiraceae* and *Gemm 1*. We also assessed the network of correlations across tumors with mutations in PID pathways (Fig. 4B). This analysis highlighted that some pathway-level mutations show a shared set of correlations between taxa, while others appear independent. Several of the taxa that can be used to predict LoF mutations in p75(NTR) signaling share correlations among each other as well as with taxa associated with mutations in PDGFR-beta signaling and direct p53 effectors.

Discussion

The link between colorectal cancer and the gut microbiome has been highlighted by a large number of recent studies [2–18], with several hypotheses as to the causal role of microbes in the disease [9,12,36,37]. Since cancer is a genetic disease caused by mutations in host DNA, it is of interest to study the microbiome in the context of tumor mutational profiles, especially given recent studies showing an impact of host genetics on the microbiome[24–29]. Here, we jointly

analyzed tumor coding mutational profile and the taxonomic composition of the proximal microbiome. We found that the composition of the proximal microbiome is correlated with mutations in tumor DNA, and that this correlation can be used to predict mutated genes and pathways solely based on the microbiome.

We performed quality control of the data and stringent filtering at every step (*e.g.*, requiring 30x coverage at a site in both the tumor and matched normal sample to call a mutation; see methods). While these requirements are likely to increase the frequency of false negatives (true mutations that simply do not meet our criteria), this rigorous strategy is appropriate as a means of increasing the biological relevance of our findings. Of note, when comparing the common LoF mutations found in our dataset to those found in colorectal tumors sampled as part of The Cancer Genome Atlas (TCGA) project, we find several commonalities, including a high frequency of LoF mutations in *APC,* as well as numerous missense mutations in *KRAS*, *NRAS,* and *TP53*, as expected (Supplementary Table 1)[38]. In general, the numbers of mutations across our sample set were also in line with those identified at part of the TCGA (Supplementary Table 2)[38]

The association of microbial taxa with tumor stage (Fig. 1) mirrors recent results, including a study of a Chinese population[4,35]. This concordance is relevant as it indicates that the microbial communities appear to be consistent even when comparing geographically distinct patient cohorts[39,40]. One of the predictive taxa, *Porphyromonadaceae*, has been shown to be altered in mouse models of CRC in other studies as well[7,14]. A study on the link between dysbiosis and colitis-induced colorectal cancer also showed similar results[41]. For instance, the bacterial genus *Paludibacter* was found to be associated with risk of developing tumors in a mouse model[41]. We find that *Paludibacter* is significantly associated with low-stage tumors, again, supporting the hypothesis that these bacteria as associated with cancer risk and may be contributing to early stage inflammation[41]. Conversely, we found that the genus *Coprococcus* is associated with high-stage tumors and not low stage tumors. Members of this genus are known to generate butyrate and propionate, which in this context can act as antiinflammatory short chain fatty acids[42]. Although our results are correlational and cannot point to causal effects, these findings suggest that driving inflammation may play a role in early stage cancer, while generating nutrients at the cost of suppressing inflammation may be more beneficial to the tumor in later stages.

Gene-level mutation data, visualized in Supplementary Fig. 3, show intriguing patterns of microbial abundances that are associated with the tumors harboring different mutations. For instance, as reflected in the differing patterns within each gene (rows) in the heatmap, *Aerococcus* and *Dorea* are both show higher scaled abundances within tumors harboring mutations in *ZNF717*, *CTBP2*, and *APC*, relative to tumors with LoF mutations in *ANKRD36C* and *KMT2C*. This highlights the different patterns in the microbiome that can be found when assessing genetically heterogeneous sets of tumors; as *Dorea* has been found to be increased in tumor microbiomes by several different groups, whereas our work highlights some potential genetic interactions that explain cases wherein *Dorea* is not increased at the tumor site[3,5–8]. Thus, incorporating genetic profiles in studies of the microbiome in CRC may be beneficial and uncover patterns that are dependant on specific tumor mutations.

Although it may be difficult to ascertain the biological mechanism behind the predicted interactions among mutated genes and microbial taxa (shown in Fig. 2), it is possible to generate hypotheses based on what is already known in the relevant literature. For example, *ANKRD36C* encodes a protein that may have a role in ion transport in epithelial cells[43]. Additionally, we find that LoF mutations in *APC* correlate with changes in 25 different microbial taxa, including an increase in the abundance of the genus *Finegoldia*. This genus has been identified in previous studies of colon adenomas and also harbors species that act as opportunistic pathogens at sites where the epithelium has been damaged[6,44,45]. In addition, *Capnocytophaga* has been identified as a potential biomarker for lung cancer[46]. Interestingly, changes in the abundance of *Christensenellaceae* are predictive of mutations in both *APC* and *ZNF717*. A recent study in twins has identified *Christensenellaceae* as a taxon that is highly driven by host genetics[26]. We find that mutations in *ZNF717*, a transcription factor commonly altered in gastric, hepatocellular, and cervical cancers [47–49], are associated with *Verrucomicrobiaceae* and *Akkermansia*, which are both known to increase in conjunction with colitis[50]. *Alphaproteobacteria* are significant contributors to our ability to predict mutations in *CTBP2*, a repressor of transcription known to interact with the ARF tumor suppressor[51]. Changes in this bacterial taxon's abundance has also been found to be associated with prostate cancer, however a mechanism of action was not explored[52]. We also show that mutations in *KMT2C*, a gene commonly co-mutated along with *KRAS,* could be predicted, in part, using the abundance of *Ruminococcus*[53]. These bacteria have been previously implicated in inflammatory bowel disorders and colorectal cancer by multiple groups[8,54–56].

Similar results were also evident when aggregating the mutations into KEGG and PID pathways (Fig. 3, Supplementary Figs. 4-5; see Methods) [30–33]. As an example, we find that the abundance of microbes that predict KEGG pathways form two distinct clusters, and that the genus *Escherichia* has a higher scaled abundance in tumors with mutations in the KEGG pathways in cluster 1 relative to those in cluster 2 (Supplementary Fig. 4). Cluster 1 contains adherens junctions, which are partially responsible for maintaining the intestinal barrier and interestingly, a disruption of the intestinal barrier in mice using cyclophosphamide was shown to cause a loss of adherens junction function and a concomitant increase in bacterial translocation into the intestinal tissue, including species of *Escherichia*[57]. When examining the heatmap with LoF mutation collapsed into PID pathways (Supplementary Fig. 5), we again find differences in scaled microbial abundances between the tumors as a function of which pathways are mutated. For instance, we find lower abundance of *Pseudomonas* in tumors with LoF mutations in the pathways 'regulation of nuclear β-catenin signaling and target gene transcription', 'degradation of β-catenin', 'presenilin action in Notch and Wnt signaling', and 'canonical Wnt signaling pathways'. Recent studies have shown that *Pseudomonas* strains that express the *LecB* gene can lead to degradation of β-catenin, providing hypothetical support for the concept that this genus may play a somewhat protective role in CRC by suppressing the Wnt signaling pathway[58]. The mechanism that might explain this phenomenon is still unclear but may have to do with alterations in appropriate cell surface adhesion molecules for the LecB protein or a change in the content of the cellular microenvironment[58,59].

Many of the interactions identified here between bacterial taxa and mutations in PID pathways have been demonstrated experimentally in the literature. For example, in human oral

cancer cells, it was shown that bacteria of interest were able to activate EGFR through the generation of hydrogen peroxide[60]. In addition, the correlation between ErbB1 downstream signaling and increase in the abundance of *Corynebacterium* has been demonstrated mechanistically in a model of atopic dermatitis, whereby EGFR inhibition results in dysbiosis (the appearance of *Corynebacterium* species) and inflammation[61]. Specific depletion of *Corynebacterium* ablates the inflammatory response[61]. Moreover, our finding that the abundance of *Fusobacterium* is depleted in tumors with LoF mutations in the PDGFR-beta pathway, which may be explained by the dependence of several pathogenic strains of bacteria for functionally intact PDGFR signaling for adherence to intestinal epithelium[62]. In addition, p75(NTR) signaling has been shown to operate as a tumor suppressor by mediating apoptosis in response to hypoxic conditions and reactive oxygen species[63–66]. Alterations in this pathway have also been shown to be useful as a biomarker for esophageal cancer[67,68].

Our study has several caveats. First, our study only shows correlations, and we cannot directly assess causal effects. Thus, we do not know whether the microbiome is altered before or after the appearance of specific mutations. Nevertheless, many of the predicted interactions described above have been previously tested, albeit across a wide variety of experimental systems and disease states, typically in isolation, for biological relevance and mechanism of action. We expect that future studies will more comprehensively test the causality of interactions by utilizing model organisms and cell culture techniques, where the effect of individual mutations can be assessed. Additionally, we have only profiled the taxonomic composition of the microbiome, and thus cannot detect interactions that are dependent on microbial genes or functions. Similarly, using whole-exome sequencing does not allow us to include non-coding mutations and larger tumor structural variants and chromosomal abnormalities. This can be alleviated by the use of metagenomic shotgun sequencing to profile the microbiome, as well as whole-genome sequencing to assess tumor mutations. Moreover, the study sample was relatively small (n = 88 samples from 44 patients). Nevertheless, the sample size was sufficient to detect significant patterns. Additional studies that use large tumor samples would be useful in validating our results and identifying further associations.

In summary, we present a strong association between tumor genetic profiles and the proximal microbiome, and identify tumor genes and pathways that correlate with specific microbial taxa. We also show that the microbiome can be used as a predictor of tumor mutated genes and pathways, and suggest potential mechanisms driving the interaction between the tumor and its microbiota. Our proof-of-principle analysis can provide a starting point for the development of diagnostics that utilize microbiome profiles to ascertain CRC tumor mutational profiles, facilitating personalized treatments.

Methods

Patient inclusion and DNA extraction

88 tissue samples from 44 individuals were used, with one tumor and one normal sample from each individual. These de-identified samples were obtained from the University of Minnesota Biological Materials Procurement Network (Bionet), a facility that archives research

samples from patients who have provided written, informed consent. These samples were previously utilized and are described in detail in a previous study[69]. To reiterate these points, all research conformed to the Helsinki Declaration and was approved by the University of Minnesota Institutional Review Board, protocol 1310E44403. Tissue pairs were resected concurrently, rinsed with sterile water, flash frozen in liquid nitrogen, and characterized by staff pathologists. The criteria for selection were limited to the availability of patient-matched normal and tumor tissue specimens. Additional patient metadata are provided in the indicated work[69].

Microbiome characterization

The microbiome data used in the study was generated previously and is described exhaustively in[69]. Briefly, microbial DNA was extracted from patient-matched normal and tumor tissue samples using sonication for lysis and the AllPrep nucleic acid extraction kit (Qiagen, Valencia, CA). The V5-V6 regions of the 16S rRNA gene were PCR amplified with the addition of barcodes for multiplexing using the forward and reverse primer sets V5F and V6R from Cai, et al.[70]. The barcoded amplicons were pooled and Illumina adapters were ligated to the reads. A single lane on an Illumina MiSeq instrument was used (250 cycles, paired-end) to generate 16S rRNA gene sequences. The sequencing resulted in approximately 10.7 million total reads passing quality filtering in total, with a mean value of 121,470 quality reads per sample. The forward and reverse read pairs were merged using the USEARCH v7 program 'fastq_mergepairs', allowing stagger, with no mismatches allowed[71]. OTUs were picked using the closed-reference picking script in QIIME v1.7.0 using the Greengenes database (August 2013 release)[72–74]. The similarity threshold was set at 97%, reverse-read matching was enabled, and reference-based chimera calling was disabled.

Exome sequence data generation

Genomic DNA samples were quantified using a fluorometric assay, the Quant-iT PicoGreen dsDNA Assay Kit (Life Technologies, Grand Island, NY). Samples were considered passing quality control (QC) if they contained greater than 300 nanograms (ng) of DNA and display an A260:280 ratio above 1.7. Full workflow details for library preparation are outlined in the Nextera Rapid Capture Enrichment Protocol Guide (Illumina, Inc., San Diego, CA). In brief, libraries for Illumina next-generation sequencing were generated using Nextera library creation reagents (Illumina, Inc., San Diego, CA). A total of 50 ng of genomic DNA per sample were used as input for the library preparation. The DNA was tagmented (simultaneously tagged and fragmented) using Nextera transposome based fragmentation and transposition as part of the Nextera Rapid Capture Enrichment kit (Illumina, Inc., San Diego, CA). This process added Nextera adapters with complementarity to PCR primers containing sequences that allow addition of Illumina flow cell adapters and dual-indexed barcodes. The tagmented DNA was amplified using dual indexed barcoded primers. The amplified and indexed samples were pooled (8 samples per pool) and quantified to ensure appropriate DNA concentrations and fragment sizes using the fluorometric PicoGreen assay and the Bioanalyzer High-Sensitivity DNA Chip (Agilent Technologies, Santa Clara, CA). Libraries were considered to pass QC as long as they contained more than 500 ng of DNA and had an average peak size between 200 - 1000 base pairs. For hybridization and sequence capture, 500 nanograms of amplified library was hybridized to biotinylated oligonucleotide probes complementary to regions of interest at 58° C

for 24 hours. Library-probe hybrids were captured using streptavidin-coated magnetic beads and subjected to multiple washing steps to remove non-specifically bound material. The washed and eluted library was subjected to a second hybridization and capture to further enrich target sequences. The captured material was then amplified using 12 cycles of PCR. The captured, amplified libraries underwent QC using a Bioanalyzer, and fluorometric PicoGreen assay. Libraries were considered to pass QC as long as they contained a DNA concentration greater than 10 nM and had an average size between 300 - 400 base pairs. Libraries were hybridized to a paired end flow cell at a concentration of 10 pM and individual fragments were clonally amplified by bridge amplification on the Illumina cBot (Illumina, Inc., San Diego, CA). Eleven lanes on an Illumina HiSeq 2000 (Illumina, Inc., San Diego, CA) were required to generate the desired sequences. Once clustering was complete, the flow cell was loaded on the HiSeq 2000 and sequenced using Illumina's SBS chemistry at 100 bp per read. Upon completion of read 1, base pair index reads were performed to uniquely identify clustered libraries. Finally, the library fragments were resynthesized in the reverse direction and sequenced from the opposite end of the read 1 fragment, thus producing the paired end read 2. Full workflow details are outlined in Illumina's cBot User Guide and HiSeq 2000 User Guides. Base call (.bcl) files for each cycle of sequencing were generated by Illumina Real Time Analysis (RTA) software. The base call files and run folders were then exported to servers maintained at the Minnesota Supercomputing Institute. Primary analysis and de-multiplexing was performed using Illumina's CASAVA software 1.8.2. The end result of the CASAVA workflow was de-multiplexed FASTQ files that were utilized in subsequent analysis for read QC, mapping, and mutation calling.

Exome data analysis

The exome sequence data contained approximately 4.2 billion reads in total following adapter removal and quality filtering, inclusive of forward and reverse reads, with a mean value of 47.8 million high-quality reads per sample. The raw reads were assessed using FastQC v0.11.2 and the Nextera adapters removed using cutadapt v1.8.1[75,76]. Simultaneously, cutadapt was used to trim reads at bases with quality scores less than 20. Reads shorter than 40 bases were excluded. The trimmed and filtered read pairs were aligned and mapped to the human reference genome (hg19) using bwa v0.7.10 resulting in a bam file for each patient sample[77]. These files were further processed to sort the reads, add read groups, correct the mate-pair information, and mark and remove PCR duplicates using picard tools v1.133 and samtools v0.1.18[78,79]. Tumor-specific mutations were identified using FreeBayes v0.9.14-24-gc292036[80]. Following these steps, 94.0% of the remaining read pairs mapped to the reference genome, hg19. Specifically, SNPs only were assessed and a minimum coverage at each identified mutation position of more than 30X was required in both the patient normal and tumor samples. These mutations were filtered to only include those that were within protein-coding regions and compiled into a single vcf file. This vcf file was assessed using SNPeffect v4.1 K (2015-09-0) in order to predict the potential impact of each of the mutations[81]. Based on these results, the mutations were grouped into three categories: (1) total mutations (2) non-synonymous mutations and (3) loss of function (LoF) mutations. The total mutations group is self-explanatory. The non-synonymous mutations included all the mutations in the total mutations group that were non-silent. The LoF group only included those mutations that resulted in a premature stop codon, a loss of a stop codon, or a frameshift mutation.

Joint analysis of microbiome and exome data

Taxa that differentiated patients with or without LoF mutation were identified using LEfSe[34]. All the taxa with a LDA score (log 10) > 2 were included in the calculation of the risk indices, built to predict the presence or absence of a LoF mutation based on the OTU table collapsed at genus level. To build the risk index, the relative abundances (arcsine square root transformed) of the taxa associated with the LoF mutation (based on the LEfSe output) were summed and the relative abundances of the taxa associated with no mutation (based on the LEfSe output) were summed. The use of the unweighted sum in the risk index, rather than relying on the regression coefficients from LDA, is a simple way to control the degree of flexibility of the model when training on small sample sizes. More detail is described in a previous publication[82]. Then the difference between these two sums was calculated, thereby obtaining a risk index. This procedure was repeated 44 times to obtain a risk index for each patient.

A leave-one-out procedure (also described above) was conducted to evaluate the taxa that differentiated patients with or without LoF mutation in the held-out patient, based on the LEfSe output of n-1 patients. In detail, the taxa that differentiated patients with or without LoF mutation were identified using LEfSe in the n-1 dataset. The relative abundances of the taxa associated with the LoF mutation (based on the LEfSe output of the n-1 dataset) were summed and the relative abundances of the taxa associated with no mutation (based on the LEfSe output of the n-1 dataset) were summed and were used to build the risk index in the held-out patient. In detail, the difference between these two sums was calculated to obtain the risk index of the held-out patient. This procedure was repeated 44 times, to produce a risk index in each of the held-out patients, based on the difference between the sum of the taxa associated with the absence of LoF mutation minus the sum of the taxa associated with the presence of the LoF mutation found in each of the n-1 datasets. The significance of the difference in risk indexes between the patients with LoF mutation and patients with LoF mutation for each gene was assessed using a Mann-Whitney U test and a permutation test, in which we permuted the labels for a given gene 999 times, each time deriving new held-out predictions of the risk indexes for each subject for that gene. Then the observed difference in means between the patients with LoF mutation and patients with LoF mutation risk index predictions using the method on the actual LoF mutation labels to the differences observed in the permutations to obtain an empirical P-value was compared. The resulting P-values were corrected using the false discovery rate (FDR) correction for multiple hypothesis tests.

Receiving Operating Characteristic (ROC) curves were plotted and the area under the curve (AUC) values computed on a dataset containing 10 sets of predictions and corresponding labels obtained from 10-fold cross-validation using ROCR package in R[83]. A risk index threshold was also obtained that best predicts the presence or absence of LoF mutation with a leave-one-out cross-validation on the risk index. Each held-out sample was treated as a new patient on whom the optimal risk index cutoff was tested and subsequently refined to separate patients who had a LoF mutation and patient who did not have a LoF.

Correlation analysis was performed using SparCC on a reduced OTU table containing significant taxa identified using the above prediction methods collapsed to the genus level[84].

Pseudo p-values were calculated using 100 randomized sets. Networks of correlations were visualized using Cytoscape v3.1.0[85].

As this work is an extension of a previous study of the CRC-associated microbiome, each of the patients in this project have associated clinical data[69]. We used a linear model to determine the extent to which any of these factors may correlate with mutation load. These included patient sex, tumor stage, patient age, patient body mass index (BMI), and microsatellite instability (MSI) status. None of these factors, alone or in combination, were found to significantly impact the mutational data, though it bears noting that MSI status was only available for a subset (13 out of 44) of the patients.

Acknowledgments

We thank the members of the Blekhman Lab for helpful discussions. The microbiome sequencing data reported here can be accessed at the NCBI Sequence Read Archive under project accession PRJNA284355. The exome sequencing data described here can be accessed from dbGAP under reference number [yet to be determined - submission in process]. This work is supported in part by funds from the University of Minnesota College of Biological Sciences (R.B.), The Randy Shaver Cancer Research and Community Fund (R.B), Institutional Research Grant #124166-IRG-58-001-55-IRG53 from the American Cancer Society (R.B.), a Research Fellowship from The Alfred P. Sloan Foundation (R.B.), and a Translational Research Development Grant from the University of Minnesota Clinical and Translational Science Institute (M.B.B). This work was supported, in part, by resources provided by the Minnesota Supercomputing Institute.

Author Contributions

M.B.B. and R.B. designed the project. M.B.B. performed DNA extractions, exome sequencing analysis, microbiome sequencing analysis, prepared the manuscript, and generated the figures. E.M. and M.B.B performed statistical analyses including LEfSe, the LOO approach, ROC curve generation, and permutation tests. J.A. assisted with the exome sequencing data analysis pipeline. T.K.S. contributed to the experimental design and interpretations of data. D.K. contributed to the design and interpretation of the interaction prediction approach. R.B. provided guidance related to the analytical approaches and interpretation of the results. All authors contributed to manuscript and figure revisions.

Competing Financial Interests

The authors declare no competing financial interests.

References

1. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486,** 207–214 (2012).
2. Burns, M., Lynch, J., Starr, T., Knights, D. & Blekhman, R. Virulence genes are a signature of the microbiome in the colorectal tumor microenvironment. *Genome Med.* **7,** 55 (2015).
3. Warren, R. L. *et al.* Co-occurrence of anaerobic bacteria in colorectal carcinomas. *Microbiome* **1,** 16 (2013).
4. Nakatsu, G. *et al.* Gut mucosal microbiome across stages of colorectal carcinogenesis. *Nat. Commun.* **6,** 8727 (2015).
5. Wang, T. *et al.* Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *ISME J.* **6,** 320–329 (2012).
6. Shen, X. J. *et al.* Molecular characterization of mucosal adherent bacteria and associations with colorectal adenomas. *Gut Microbes* **1,** 138–147 (2010).
7. Chen, W., Liu, F., Ling, Z., Tong, X. & Xiang, C. Human intestinal lumen and mucosa-associated microbiota in patients with colorectal cancer. *PLoS One* **7,** e39743 (2012).
8. Zhu, Q. *et al.* Analysis of the intestinal lumen microbiota in an animal model of colorectal cancer. *PLoS One* **9,** e90849 (2014).
9. Song, X. *et al.* Alterations in the microbiota drive interleukin-17C production from intestinal epithelial cells to promote tumorigenesis. *Immunity* **40,** 140–152 (2014).
10. Flemer, B. *et al.* Tumour-associated and non-tumour-associated microbiota in colorectal cancer. *Gut* (2016). doi:10.1136/gutjnl-2015-309595
11. Wynendaele, E. *et al.* Crosstalk between the microbiome and cancer cells by quorum sensing peptides. *Peptides* **64,** 40–48 (2015).
12. Rubinstein, M. R. *et al.* Fusobacterium nucleatum promotes colorectal carcinogenesis by modulating E-cadherin/β-catenin signaling via its FadA adhesin. *Cell Host Microbe* **14,** 195–206 (2013).
13. Zeller, G. *et al.* Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10,** 766 (2014).
14. Zackular, J. P., Rogers, M. A. M., Ruffin, M. T., 4th & Schloss, P. D. The human gut microbiome as a screening tool for colorectal cancer. *Cancer Prev. Res.* **7,** 1112–1121 (2014).
15. Yu, J. *et al.* Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* (2015). doi:10.1136/gutjnl-2015-309800
16. Baxter, N. T., Ruffin, M. T., 4th, Rogers, M. A. M. & Schloss, P. D. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genome Med.* **8,** 37 (2016).
17. Mima, K. *et al.* Fusobacterium nucleatum in colorectal carcinoma tissue and patient prognosis. *Gut* (2015). doi:10.1136/gutjnl-2015-310101
18. Kostic, A. D. *et al.* Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome Res.* **22,** 292–298 (2012).
19. Arthur, J. C. *et al.* Intestinal inflammation targets cancer-inducing activity of the microbiota. *Science* **338,** 120–123 (2012).
20. Arthur, J. C. *et al.* Microbial genomic analysis reveals the essential role of inflammation in

bacteria-induced colorectal cancer. *Nat. Commun.* **5,** 4724 (2014).
21. Housseau, F. & Sears, C. L. Enterotoxigenic Bacteroides fragilis (ETBF)-mediated colitis in Min (Apc+/-) mice: a human commensal-based murine model of colon carcinogenesis. *Cell Cycle* **9,** 3–5 (2010).
22. Goodwin, A. C. *et al.* Polyamine catabolism contributes to enterotoxigenic Bacteroides fragilis-induced colon tumorigenesis. *Proc. Natl. Acad. Sci. U. S. A.* **108,** 15354–15359 (2011).
23. Boleij, A. *et al.* The Bacteroides fragilis toxin gene is prevalent in the colon mucosa of colorectal cancer patients. *Clin. Infect. Dis.* **60,** 208–215 (2015).
24. Goodrich, J. K. *et al.* Genetic Determinants of the Gut Microbiome in UK Twins. *Cell Host Microbe* **19,** 731–743 (2016).
25. Blekhman, R. *et al.* Host genetic variation impacts microbiome composition across human body sites. *Genome Biol.* **16,** 191 (2015).
26. Goodrich, J. K. *et al.* Human genetics shape the gut microbiome. *Cell* **159,** 789–799 (2014).
27. Knights, D. *et al.* Complex host genetics influence the microbiome in inflammatory bowel disease. *Genome Med.* **6,** 107 (2014).
28. Davenport, E. R. *et al.* Genome-Wide Association Studies of the Human Gut Microbiota. *PLoS One* **10,** e0140301 (2015).
29. Goodrich, J. K., Davenport, E. R., Waters, J. L., Clark, A. G. & Ley, R. E. Cross-species comparisons of host genetic associations with the microbiome. *Science* **352,** 532–535 (2016).
30. Kanehisa, M. & Goto, S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28,** 27–30 (2000).
31. Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42,** D199–205 (2014).
32. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* (2015). doi:10.1093/nar/gkv1070
33. Schaefer, C. F. *et al.* PID: the Pathway Interaction Database. *Nucleic Acids Res.* **37,** D674–9 (2009).
34. Segata, N. *et al.* Metagenomic biomarker discovery and explanation. *Genome Biol.* **12,** R60 (2011).
35. Ahn, J. *et al.* Human gut microbiome and risk for colorectal cancer. *J. Natl. Cancer Inst.* **105,** 1907–1911 (2013).
36. Dennis, K. L. *et al.* Adenomatous polyps are driven by microbe-instigated focal inflammation and are controlled by IL-10-producing T cells. *Cancer Res.* **73,** 5905–5913 (2013).
37. Irrazábal, T., Belcheva, A., Girardin, S. E., Martin, A. & Philpott, D. J. The multifaceted role of the intestinal microbiota in colon cancer. *Mol. Cell* **54,** 309–320 (2014).
38. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487,** 330–337 (2012).
39. Yatsunenko, T. *et al.* Human gut microbiome viewed across age and geography. *Nature* **486,** 222–227 (2012).
40. Allali, I. *et al.* Gut microbiome compositional and functional differences between tumor and non-tumor adjacent tissues from cohorts from the US and Spain. *Gut Microbes* **6,** 161–172

(2015).
41. Couturier-Maillard, A. *et al.* NOD2-mediated dysbiosis predisposes mice to transmissible colitis and colorectal cancer. *J. Clin. Invest.* **123,** 700–711 (2013).
42. Louis, P., Hold, G. L. & Flint, H. J. The gut microbiota, bacterial metabolites and colorectal cancer. *Nat. Rev. Microbiol.* **12,** 661–672 (2014).
43. Kumar, R., Haugen, J. D., Wieben, E. D., Londowski, J. M. & Cai, Q. Inhibitors of renal epithelial phosphate transport in tumor-induced osteomalacia and uremia. *Proc. Assoc. Am. Physicians* **107,** 296–305 (1995).
44. Dowd, S. E. *et al.* Survey of bacterial diversity in chronic wounds using pyrosequencing, DGGE, and full ribosome shotgun sequencing. *BMC Microbiol.* **8,** 43 (2008).
45. Murphy, E. C. *et al.* Identification of molecular mechanisms used by Finegoldia magna to penetrate and colonize human skin. *Mol. Microbiol.* **94,** 403–417 (2014).
46. Yan, X. *et al.* Discovery and validation of potential bacterial biomarkers for lung cancer. *Am. J. Cancer Res.* **5,** 3111–3122 (2015).
47. Cui, J. *et al.* Comprehensive characterization of the genomic alterations in human gastric cancer. *Int. J. Cancer* **137,** 86–95 (2015).
48. Chen, Y., Wang, L., Xu, H., Liu, X. & Zhao, Y. Exome capture sequencing reveals new insights into hepatitis B virus-induced hepatocellular carcinoma at the early stage of tumorigenesis. *Oncol. Rep.* **30,** 1906–1912 (2013).
49. Lando, M. *et al.* Interplay between promoter methylation and chromosomal loss in gene silencing at 3p11-p14 in cervical cancer. *Epigenetics* **10,** 970–980 (2015).
50. Berry, D. *et al.* Phylotype-level 16S rRNA analysis reveals new bacterial indicators of health state in acute murine colitis. *ISME J.* **6,** 2091–2106 (2012).
51. Paliwal, S. *et al.* The alternative reading frame tumor suppressor antagonizes hypoxia-induced cancer cell migration via interaction with the COOH-terminal binding protein corepressor. *Cancer Res.* **67,** 9322–9329 (2007).
52. Yu, H. *et al.* Urinary microbiota in patients with prostate cancer and benign prostatic hyperplasia. *Arch. Med. Sci.* **11,** 385–394 (2015).
53. Fang, B. RAS signaling and anti-RAS therapy: lessons learned from genetically engineered mouse models, human cancer cells, and patient-related studies. *Acta Biochim. Biophys. Sin.* **48,** 27–38 (2016).
54. Hu, Y. *et al.* Manipulation of the gut microbiota using resistant starch is associated with protection against colitis-associated colorectal cancer in rats. *Carcinogenesis* (2016). doi:10.1093/carcin/bgw019
55. Tsuruya, A. *et al.* Major Anaerobic Bacteria Responsible for the Production of Carcinogenic Acetaldehyde from Ethanol in the Colon and Rectum. *Alcohol Alcohol* (2016). doi:10.1093/alcalc/agv135
56. Zhang, M. *et al.* Effects of Lactobacillus salivarius Ren on cancer prevention and intestinal microbiota in 1, 2-dimethylhydrazine-induced rat model. *J. Microbiol.* **53,** 398–405 (2015).
57. Yang, J., Liu, K.-X., Qu, J.-M. & Wang, X.-D. The changes induced by cyclophosphamide in intestinal barrier and microflora in mice. *Eur. J. Pharmacol.* **714,** 120–124 (2013).
58. Cott, C. *et al.* Pseudomonas aeruginosa lectin LecB inhibits tissue repair processes by triggering β-catenin degradation. *Biochim. Biophys. Acta* (2016). doi:10.1016/j.bbamcr.2016.02.004
59. Garber, N., Guempel, U., Gilboa-Garber, N. & Doyle, R. J. Specificity of the

fucose-binding lectin of Pseudomonas aeruginosa. *FEMS Microbiol. Lett.* **48,** 331–334 (1987).
60. Boonanantanasarn, K. *et al.* Enterococcus faecalis enhances cell proliferation through hydrogen peroxide-mediated epidermal growth factor receptor activation. *Infect. Immun.* **80,** 3545–3558 (2012).
61. Kobayashi, T. *et al.* Dysbiosis and Staphylococcus aureus Colonization Drives Inflammation in Atopic Dermatitis. *Immunity* **42,** 756–766 (2015).
62. Manthey, C. F. *et al.* Indispensable functions of ABL and PDGF receptor kinases in epithelial adherence of attaching/effacing pathogens under physiological conditions. *Am. J. Physiol. Cell Physiol.* **307,** C180–9 (2014).
63. Pflug, B. R., Onoda, M., Lynch, J. H. & Djakiew, D. Reduced expression of the low affinity nerve growth factor receptor in benign and malignant human prostate tissue and loss of expression in four human metastatic prostate tumor cell lines. *Cancer Res.* **52,** 5403–5406 (1992).
64. Kraemer, B. R. *et al.* A role for the p75 neurotrophin receptor in axonal degeneration and apoptosis induced by oxidative stress. *J. Biol. Chem.* **289,** 21205–21216 (2014).
65. Wang, T.-C., Luo, S.-J., Lin, C.-L., Chang, P.-J. & Chen, M.-F. Modulation of p75 neurotrophin receptor under hypoxic conditions induces migration and invasion of C6 glioma cells. *Clin. Exp. Metastasis* **32,** 73–81 (2015).
66. Le Moan, N., Houslay, D. M., Christian, F., Houslay, M. D. & Akassoglou, K. Oxygen-dependent cleavage of the p75 neurotrophin receptor triggers stabilization of HIF-1α. *Mol. Cell* **44,** 476–490 (2011).
67. Yamaguchi, T. *et al.* Detection of circulating tumor cells by p75NTR expression in patients with esophageal cancer. *World J. Surg. Oncol.* **14,** 40 (2016).
68. Yamaguchi, T. *et al.* p75 neurotrophin receptor expression is a characteristic of the mitotically quiescent cancer stem cell population present in esophageal squamous cell carcinoma. *Int. J. Oncol.* **48,** 1943–1954 (2016).
69. Burns, M. B., Lynch, J., Starr, T. K., Knights, D. & Blekhman, R. Virulence genes are a signature of the microbiome in the colorectal tumor microenvironment. *Genome Med.* **7,** 55 (2015).
70. Cai, L., Ye, L., Tong, A. H. Y., Lok, S. & Zhang, T. Biased diversity metrics revealed by bacterial 16S pyrotags derived from different primer sets. *PLoS One* **8,** e53649 (2013).
71. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26,** 2460–2461 (2010).
72. Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* **7,** 335–336 (2010).
73. Navas-Molina, J. A. *et al.* Advancing our understanding of the human microbiome using QIIME. *Methods Enzymol.* **531,** 371–444 (2013).
74. DeSantis, T. Z. *et al.* Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl. Environ. Microbiol.* **72,** 5069–5072 (2006).
75. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. Available at: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/. (Accessed: 28th September 2015)
76. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* **17,** 10–12 (2011).

77. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
78. Broad Institute. Picard Tools - The Broad Institute. Available at: http://broadinstitute.github.io/picard/. (Accessed: 29th September 2015)
79. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
80. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv [q-bio.GN]* (2012).
81. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6,** 80–92 (2012).
82. Montassier, E. *et al.* Pretreatment gut microbiome predicts chemotherapy-related bloodstream infection. *Genome Med.* **8,** 49 (2016).
83. Sing, T., Sander, O., Beerenwinkel, N. & Lengauer, T. ROCR: visualizing classifier performance in R. *Bioinformatics* **21,** 3940–3941 (2005).
84. Friedman, J. & Alm, E. J. Inferring correlation networks from genomic survey data. *PLoS Comput. Biol.* **8,** e1002687 (2012).
85. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–2504 (2003).

Figure Legends

Figure 1 | Correlation between the microbial community at a tumor that differentiates between tumor stage. a, Low-stage (stages 1-2) and high-stage (stages 3-4) tumors can be differentiated using a risk index classifier generated from microbial abundance data (y-axis). The central black bar indicates the median, and the thin black bars represent the 25th and 75th percentiles. **b**, A receiver operating characteristic (ROC) curve was generated using a 10-fold cross-validation (blue dotted lines). The average of the 10-fold cross-validation curves is represented as a thick black line. **c**, Differences in the mean abundances of a subset of the taxa predicted to interact differentially with high-stage and low-stage tumors. This subset represents those taxa that had a mean difference in abundance of greater than 0.1%, proportionally.

Figure 2 | Commonly mutated genes show a predicted interaction with changes in the abundances of several microbial taxa. a, A heatmap of the scaled abundances values (cells) for a subset of taxa chosen as they were identified as discriminatory in each leave-one-out iteration (columns) that were found significantly associated with prevalent LoF mutations (rows). Scaled abundances are from the patients with the indicated mutations. **b**, LoF mutations in each of the indicated genes can be predicted using a risk index as a classifier (y-axis). The central black bar indicates the median, and the thin black bars represent the 25th and 75th percentiles. **c**, ROC curves were generated for each of the indicated mutations using a 10-fold cross-validation (blue dotted lines). The average of the 10-fold cross-validation curves is represented as a thick black line. **d**, Differences in the mean abundances of a subset of the taxa predicted to interact differentially with tumors with a LoF mutation relative to those without the indicated mutation. This subset represents those taxa that had a mean difference in abundance of greater than 0.1%, proportionally.

Figure 3 | Pathways harboring prevalent LoF mutations correlate with changes in the abundances of sets of microbial taxa. a, A heatmap of the scaled abundances values (cells) for a subset of taxa (columns) that are found significantly associated with KEGG pathways harboring LoF mutations (rows). Scaled abundances are from the patients with mutations in the indicated pathways. **b**, LoF mutations in each of the indicated pathways can be predicted using a risk index as a classifier (y-axis). The central black bar indicates the median, and the thin black bars represent the 2nd and 4th quartiles. **c**, ROC curves were generated for each of the indicated pathways using a 10-fold cross-validation (blue dotted lines). The average of the 10-fold cross-validation curves is represented as a thick black line. **d**, Differences in the mean abundances of a subset of the taxa predicted to interact differentially with tumors harboring mutations in the indicated pathways relative to those without a mutation. This subset represents those taxa that had a mean difference in abundance of greater than 0.1%, proportionally. **e - f**, Identically structured visualizations as in **a - d**, but for PID pathway data rather than the KEGG pathways.

Figure 4 | Interaction networks among bacteria are defined by host tumor mutations. a, SparCC analysis of the microbial abundances for taxa identified by LEfSe for APC with LoF mutations (left) and without mutation (right) produce distinct patterns of correlations (edges) between a common set of taxa (nodes). Direct correlations are indicated as red edges and inverse correlations as blue edges (SparCC R >= 0.25, P <= 0.05 for displayed edges). **b**, SparCC

analysis was run simultaneously for all taxa identified by LEfSe when predicting PID pathways. There are interactions (dashed edges) between the taxa (grey nodes) associated with mutations across sets of PID pathways (green nodes). The solid edges indicate SparCC R-values (red for direct and blue for inverse correlations). The grey taxon nodes are scaled to the average abundance of the taxa in the associated tumor set. Edge color indicates the direction of the interaction, red for negative and blue for positive.

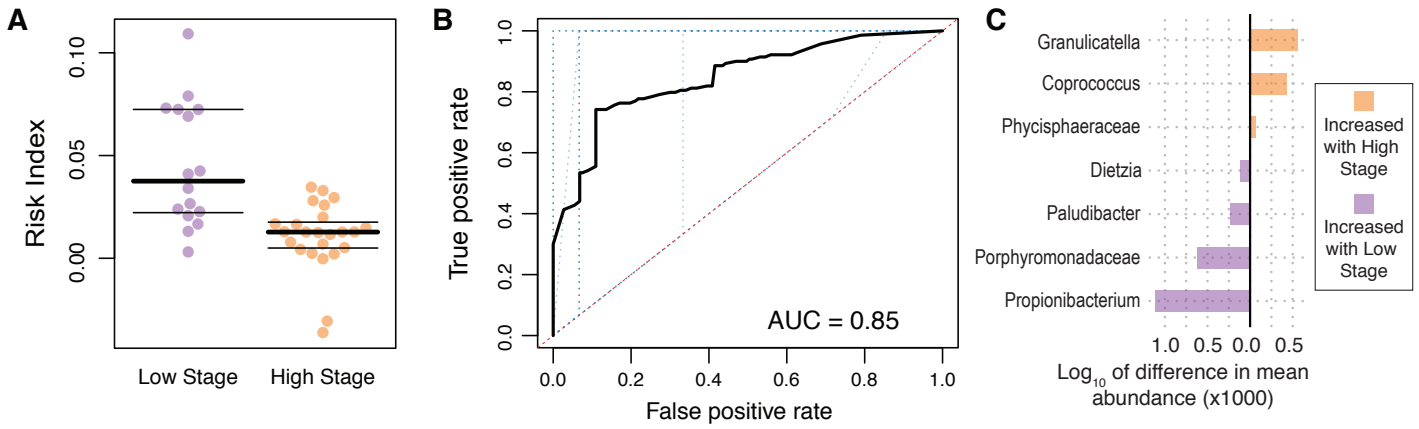
A
Risk Index
0.00
0.05
0.10
Low Stage
High Stage
B
True positive rate
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
AUC = 0.85
C
Granulicatella
Coprococcus
Phycisphaeraceae
Dietzia
Paludibacter
Porphyromonadaceae
Propionibacterium
1.0 0.5 0.0 0.5
Log10 of difference in mean abundance (x1000)
Increased with High Stage
Increased with Low Stage

Figure 1

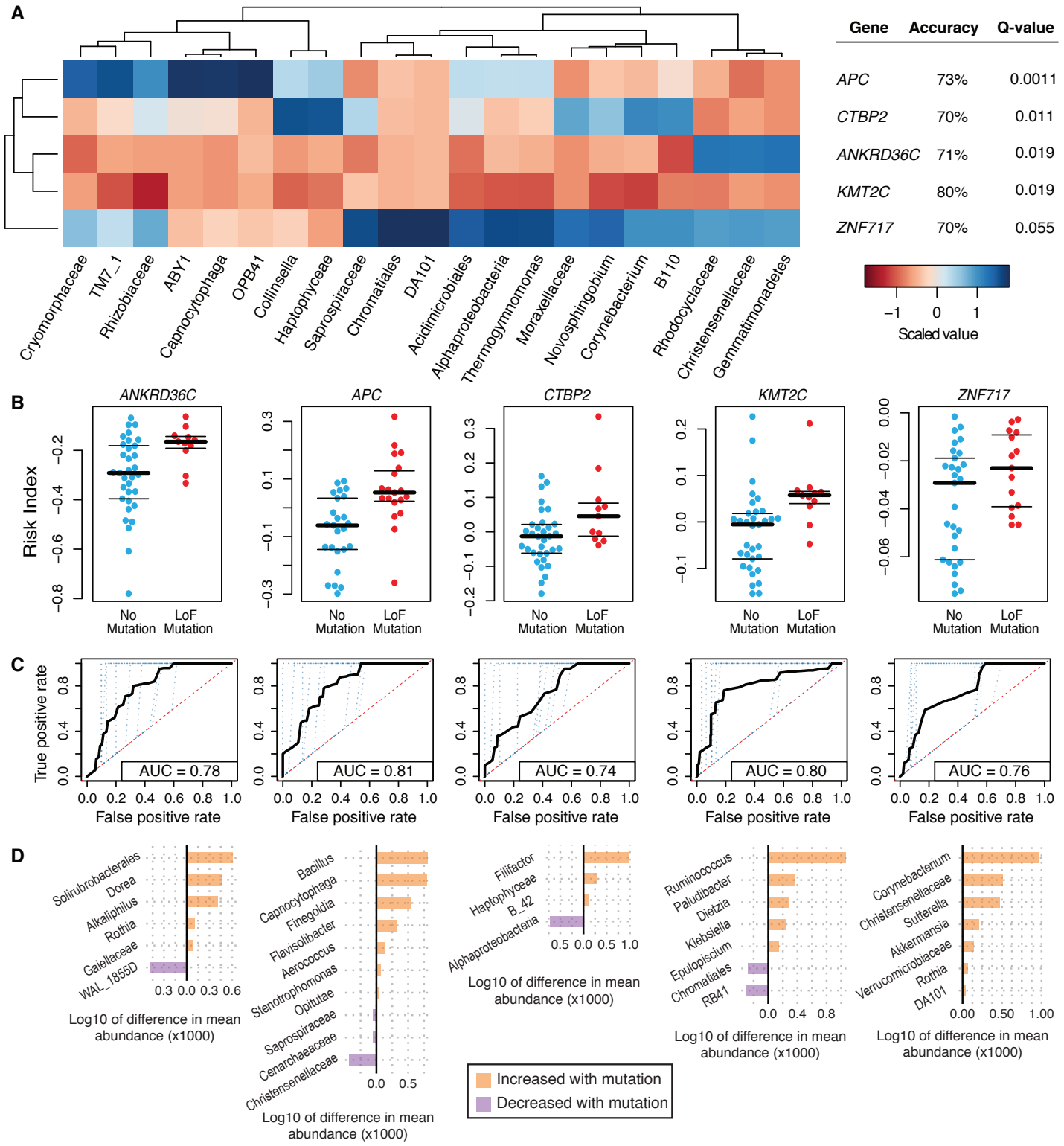
A
Gene Accuracy Q-value
APC 73% 0.0011
CTBP2 70% 0.011
ANKRD36C 71% 0.019
KMT2C 80% 0.019
ZNF717 70% 0.055
Cryomorphaceae
TM7_1
Rhizobiaceae
ABY1
Capnocytophaga
OPB41
Collinsella
Haptophyceae
Saprospiraceae
Chromatiales
DA101
Acidimicrobiales
Alphaproteobacteria
Thermogymnomonas
Moraxellaceae
Novosphingobium
Corynebacterium
B110
Rhodocyclaceae
Christensenellaceae
Gemmatimonadetes
−1 0 1
Scaled value
B
ANKRD36C
APC
CTBP2
KMT2C
ZNF717
Risk Index
No Mutation
LoF Mutation
C
True positive rate
False positive rate
AUC = 0.78
AUC = 0.81
AUC = 0.74
AUC = 0.80
AUC = 0.76
D
Solirubrobacterales
Dorea
Alkaliphilus
Rothia
Gaiellaceae
WAL_1855D
Log10 of difference in mean abundance (x1000)
Bacillus
Capnocytophaga
Finegoldia
Flavisolibacter
Aerococcus
Stenotrophomonas
Opitutae
Saprospiraceae
Cenarchaeaceae
Christensenellaceae
Filifactor
Haptophyceae
B_42
Alphaproteobacteria
Ruminococcus
Paludibacter
Dietzia
Klebsiella
Eplopiscium
Chromatiales
RB41
Corynebacterium
Christensenellaceae
Sutterella
Akkermansia
Verrucomicrobiaceae
Rothia
DA101
Increased with mutation
Decreased with mutation

Figure 2

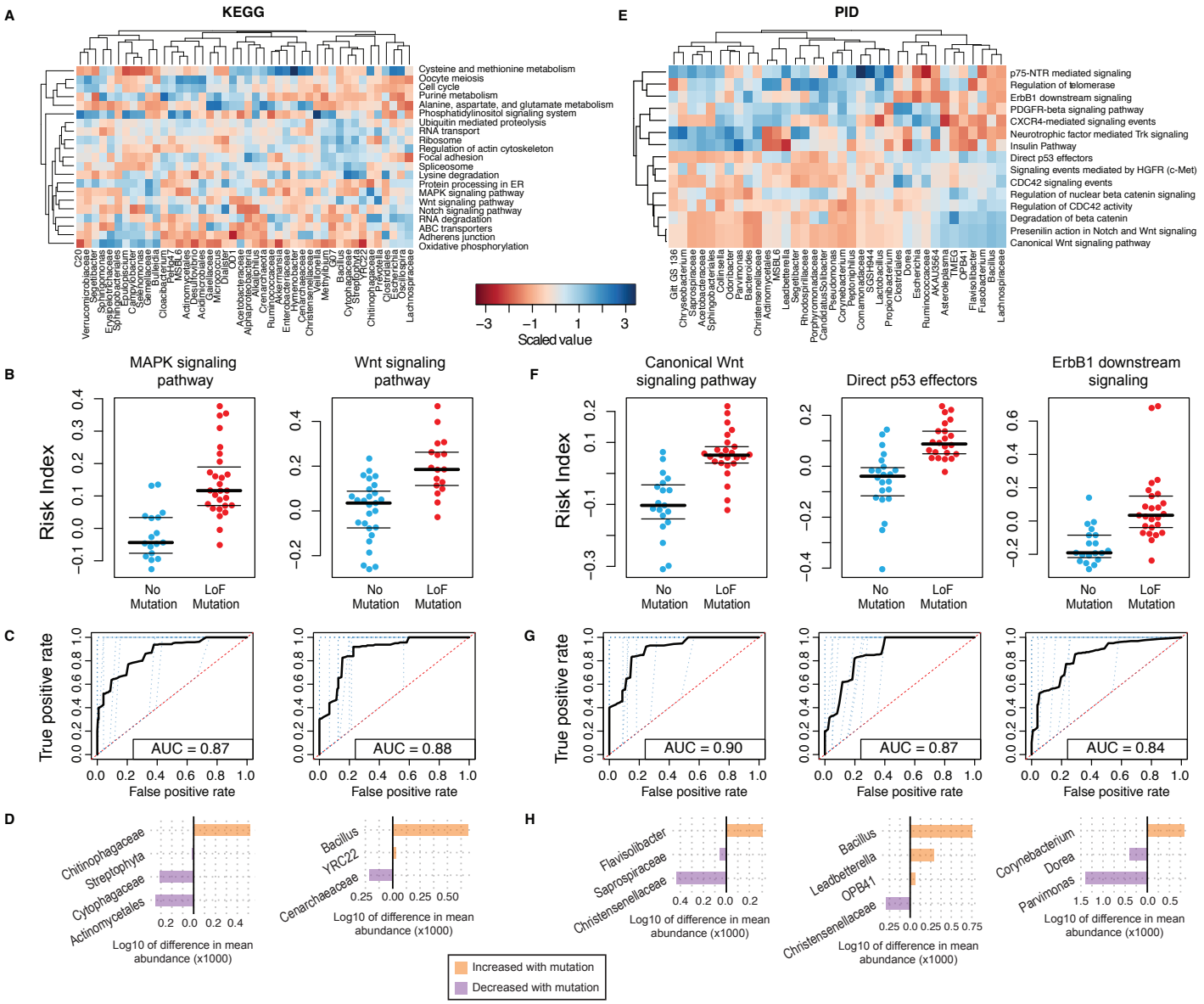
A
KEGG
Cysteine and methionine metabolism
Oocyte meiosis
Cell cycle
Purine metabolism
Alanine, aspartate, and glutamate metabolism
Phosphatidylinositol signaling system
Ubiquitin mediated proteolysis
RNA transport
Ribosome
Regulation of actin cytoskeleton
Focal adhesion
Spliceosome
Lysine degradation
Protein processing in ER
MAPK signaling pathway
Wnt signaling pathway
Notch signaling pathway
RNA degradation
ABC transporters
Adherens junction
Oxidative phosphorylation
Scaled value
E
PID
p75-NTR mediated signaling
Regulation of telomerase
ErbB1 downstream signaling
PDGFR-beta signaling pathway
CXCR4-mediated signaling events
Neurotrophic factor mediated Trk signaling
Insulin Pathway
Direct p53 effectors
Signaling events mediated by HGFR (c-Met)
CDC42 signaling events
Regulation of nuclear beta catenin signaling
Regulation of CDC42 activity
Degradation of beta catenin
Presenilin action in Notch and Wnt signaling
Canonical Wnt signaling pathway
B
MAPK signaling pathway
Wnt signaling pathway
Risk Index
No Mutation
LoF Mutation
F
Canonical Wnt signaling pathway
Direct p53 effectors
ErbB1 downstream signaling
C
True positive rate
False positive rate
AUC = 0.87
AUC = 0.88
G
AUC = 0.90
AUC = 0.87
AUC = 0.84
D
Chitinophagaceae
Streptophyta
Cytophagaceae
Actinomycetales
Bacillus
YRC22
Cenarchaeaceae
Log10 of difference in mean abundance (x1000)
H
Flavisolibacter
Saprospiraceae
Christensenellaceae
Leadbetterella
OPB41
Corynebacterium
Dorea
Parvimonas
Increased with mutation
Decreased with mutation

Figure 3

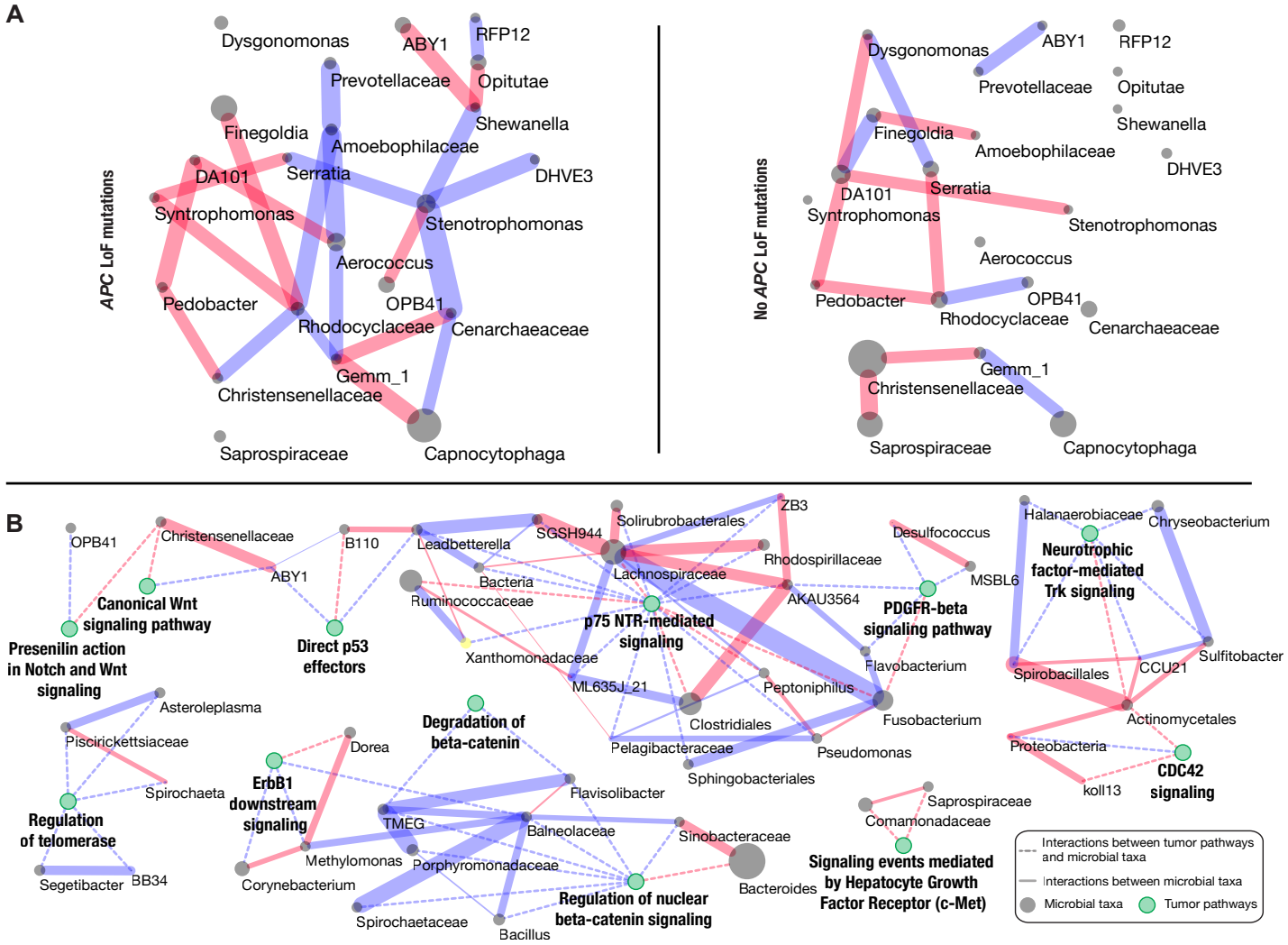
A
APC LoF mutations
Dysgonomonas
ABY1
RFP12
Prevotellaceae
Opitutae
Shewanella
Finegoldia
Amoebophilaceae
DHVE3
DA101
Serratia
Syntrophomonas
Stenotrophomonas
Aerococcus
Pedobacter
OPB41
Rhodocyclaceae
Cenarchaeaceae
Gemm_1
Christensenellaceae
Saprospiraceae
Capnocytophaga
No APC LoF mutations
Dysgonomonas
ABY1
RFP12
Prevotellaceae
Opitutae
Shewanella
Finegoldia
Amoebophilaceae
DHVE3
DA101
Serratia
Syntrophomonas
Stenotrophomonas
Aerococcus
Pedobacter
OPB41
Rhodocyclaceae
Cenarchaeaceae
Gemm_1
Christensenellaceae
Saprospiraceae
Capnocytophaga
B
OPB41
Christensenellaceae
B110
Leadbetterella
SGSH944
Solirubrobacterales
ZB3
Desulfococcus
Halanaerobiaceae
Chryseobacterium
Neurotrophic factor-mediated Trk signaling
Canonical Wnt signaling pathway
ABY1
Bacteria
Ruminococcaceae
Lachnospiraceae
Rhodospirillaceae
MSBL6
AKAU3564
PDGFR-beta signaling pathway
Presenilin action in Notch and Wnt signaling
Direct p53 effectors
p75 NTR-mediated signaling
Xanthomonadaceae
Flavobacterium
Sulfitobacter
Spirobacillales
CCU21
ML635J-21
Peptoniphilus
Asteroleplasma
Degradation of beta-catenin
Clostridiales
Fusobacterium
Actinomycetales
Piscirickettsiaceae
Dorea
Pelagibacteraceae
Pseudomonas
Proteobacteria
Sphingobacteriales
CDC42 signaling
Spirochaeta
ErbB1 downstream signaling
Flavisolibacter
koll13
Saprospiraceae
Regulation of telomerase
TMEG
Balneolaceae
Comamonadaceae
Sinobacteraceae
Methylomonas
Porphyromonadaceae
Signaling events mediated by Hepatocyte Growth Factor Receptor (c-Met)
Segetibacter
BB34
Corynebacterium
Bacteroides
Regulation of nuclear beta-catenin signaling
Spirochaetaceae
Bacillus
Interactions between tumor pathways and microbial taxa
Interactions between microbial taxa
Microbial taxa
Tumor pathways

Figure 4